AD________________

Award Number: DAMD17-01-1-0662

TITLE: Determining Effects of Genes, Environment, and Gene X Environment Interaction That are Common to Breast and Ovarian Cancers Via Bivariate Logistic Regression

PRINCIPAL INVESTIGATOR: Viswanathan Ramakrishnan, Ph.D.

CONTRACTING ORGANIZATION: Virginia Commonwealth University
Richmond, Virginia 23298-0568

REPORT DATE: July 2003

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY **(Leave blank)**	**2. REPORT DATE** July 2003	**3. REPORT TYPE AND DATES COVERED** Final (1 Jul 2001 - 30 Jun 2003)

4. TITLE AND SUBTITLE
Determining Effects of Genes, Environment, and Gene X Environment Interaction That are Common to Breast and Ovarian Cancers Via Bivariate Logisitic

5. FUNDING NUMBERS
DAMD17-01-1-0662

6. AUTHOR(S)
Viswanathan Ramakrishnan, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Virginia Commonwealth University
Richmond, Virginia 23298-0568

E-Mail: vramesh@hsc.vcu.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

A new method for the simultaneous genetic analysis of two or more discrete traits such as the presence of breast and ovarian cancers in twins was developed. A generalized estimation equations (GEE) logistic regression model was used for the modeling. A shared trait is defined for two discrete traits based upon explicit patterns of trait concordance and discordance within twin pairs; this shared trait is assessed for the influence of additive genetic and/or common environmental effects. Data are summarized in the form of 2 x 2 tables (for monozygotic and dizygotic twins) by combining appropriate cells form the 16-cell multinomial distribution to define the individual and shared trait. Hypothesis tests for additive genetic and common environmental influence are performed using repeated measures logistic regression via the GEE approach. The model specification is highly flexible, accounts for the correlated structure of the parameter estimates and does not require multivariate normality assumption for the underlying liability distribution. The approach was applied to data sets form the Vietnam Era Twin Registry and The Mid Atlantic twin Registry. Currently, efforts are being taken to collect adequate data on cancer outcomes that will provide enough power to apply this methodology to twins with cancer.

14. SUBJECT TERMS
No Subject Terms.

15. NUMBER OF PAGES
54

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

INTRODUCTION:

The research work was undertaken to develop a methodology using the logistic regression models to analyze data from two or more categorical outcomes on twins. The main objectives included a) extracting data on at least two different cancers from the Mid-Atlantic Twin Registry (MATR) sample of twins, b) Develop and apply a bivariate methodology for analyzing data on two dichotomous outcomes on the twins, such as breast and ovarian cancer and c) Develop software or macros to make the proposed method user friendly. The work has been completed in the areas of developing the statistical methods and programming. Some of this work was presented in the midterm report last year. A manuscript that has been submitted for publication is attached (Appendix A). A no cost extension was requested for an additional year to explore twin registries other than the MATR for obtaining cancer data. Although the purpose of acquiring the data was primarily to illustrate the bivariate logistic regression method (proof of concept), since bivariate analyses of cancer data have not been published in the past it was pursued also to strengthen the proposed method's usefulness. As proposed in the request for an extension, the Swedish Twin Registry was successfully contacted and the data were obtained. One of the main reasons for looking elsewhere for twin cancer data is that, unlike the Swedish Twin Registry data, the MATR cancer data are self reported data and have not been verified through diagnosis records. An attempt was made to match the Virginia cancer registry data with the MATR data to identify twins with cancer. This matching was performed using the last name, social security number (SSN) and the date of birth (Goldberg, 1993). However, this effort led to fewer than 200 twins with cancer in all. (This is partially due to the fact that the MATR data was incomplete. For example the SSN was missing for a large percent of the twins.) In this report the data obtained from the Swedish Twin Registry are summarized and the bivariate analyses performed are presented. Also, a table of articles published in the area of cancer twins is included, which will be used for future research.

Please Ignore

BODY:

1. Background

The statistical methodology for determining how genes influence the occurrence of disease is a topic of interest among both quantitative geneticists and genetic epidemiologists. Data from twin studies can be used to examine how complex genetic mechanisms and non-genetic factors influence disease occurrence. The classical twin study has long been used to estimate genetic influence on a single trait. This research design compares within-pair twin similarity for a trait in monozygotic (MZ) and dizygotic (DZ) pairs; from these similarity measures estimates of the influence of heredity and common environment are developed. The methods for the analysis of classical twin studies of continuously distributed traits (height, weight, etc) are well described and use the intra-class correlation as the principal indicator of twin similarity [Falconer and MacKay, 1996]. However, for discrete data there is no universally accepted measure of twin similarity [Hannah et al., 1983; 1985; Kendler, 1989; Neale and Cardon, 1992; Donner, 1996; Kraemer, 1997].

The odds ratio derived from logistic regression has been proposed as a measure of twin similarity in classical twin studies [Ramakrishnan et al., 1992; 1996]. The odds ratio when used with purely categorical data does not require the assumption of an underlying bivariate normal distribution. After an initial assessment of the genetic basis of a single trait it is often of interest to examine the co-aggregation of more than one trait or disease within twin pairs. For example, one might want to explore the co-occurrence of two types of cancer (breast and ovarian) or two types of substance abuse (alcohol and illicit drugs) in pairs of twins. Twins are especially useful for these types of analyses since it is possible to examine if two traits are influenced by shared additive genetic effects. However, to address the co-aggregation of two traits the univariate methods of twin analysis have to be extended to the bivariate case.

A new method for the simultaneous genetic analysis of two or more discrete traits is proposed. This method uses a GEE method to fit the multivariate logistic regression model [Liang and Zeger, 1986; Liang et al., 1992; Zeger and Liang, 1986; 1992].

2. Review and Methods

In the midterm report sent in 2002 more detailed review and method sections were presented. Please see the attached manuscript for a description of the methods (Appendix A). Work is underway to prepare a research synthesis (meta-analysis) for multiple cancers in the elderly twins (over 65). Here, a collection of articles on twin cancer research, which were examined during the last year for this purpose are tabulated (Table I) with some additional details.

Table I. Published articles in Cancer Twin Research.

Caner Site (N)	First Author	Number of Twin Pairs	Population
Prostate (1,009)	Page (1997)	15,924	White males, U.S.
Prostate, Lung, Breast	Ahlbom (1997)	N/A	Swedish Males, Swedish females
Prostate (458)	Gronberg (1994)	4,840 1649 MZ, 2,983	Swedish Twins
Prostate	Bansal (2000)	167 (84 MZ, 83 DZ)	White Male twin registry, ages 22-77
Colorectal (498)	Terry (2001)	8559	
All sites (10,803)	Lichtenstein (2000)	44,788	Swedish, Danish, Finnish
Breast	Cerhan (2000)	538 (130 MZ, 337 DZ)	Iowa - Post-menopausal female
Breast	Ekborn (1997)	N/A	Swedish
Breast	Sanderson (1996)	N/A	Seattle WA population
Breast (245)	Verkasalo (1999)	13,176 4,308 MZ, 8,868 DZ	Finnish Twin Registry
All Cancers 1,613) Prostate (202) Breast (245) Smoking-Related (1,263)	Verkasalo (1999)	25,882 8,087 MZ	Finnish Twin Registry Same Sex Twins Only
Endometrial (133)	Terry (1999)	11,659	Swedish Twin Registry
Breast (500) Testicular (194)	Swerdlow (1997)		British and Welsh young adults
Testicular (119)	Swerdlow (1999)	60	British and Welsh same sex twins
Breast	Peto (2000)		
Lung	Harris (1995)		

3. Application of the logistic regression method for the bivariate twin data.

Swedish Twin Registry Data:

The Swedish Twin Registry consists of over 140,000 twins belonging to three birth cohorts: the older cohort, the middle cohort, and the younger cohort. The first cohort of approximately 11,000 pairs, consists of same sex twins who were born between 1886 to 1925 and were alive in 1961. The second cohort of approximately 14,000 pairs consists of same sex twins who were born between 1926 through 1958 and were alive in 1972. The third cohort of approximately 22,000 pairs consists of all twins born between 1968 and 1990. For the cancer studies the older and the middle cohorts were interviewed through questionnaires and the response rates were 81% and 83%, respectively. The vital status and any diagnoses of cancer were subsequently obtained from the Swedish mortality registry and the Swedish cancer registry. The twins were matched with these registry data through a unique national registration number assigned to the Swedish citizens (similar to the Social Security Number in the U.S.). A total of 4490 cancers from the first cohort (prevalent during 1961 to 1995) and 1157 cancers from the second cohort (prevalent during 1973 to 1995) were identified.

The table II below summarizes the cancer data on these twins. For the purpose of performing the bivariate analyses the data are presented in terms of pairs of cancers. The frequencies for the last cancer pair, namely, colon and rectal cancers, include both male-male and female-female pairs (but not the male-female pairs), while the other cancer pairs include only the female-female twins. If there is an additive genetic component present for the co-aggregation (which will be called linked trait) between any two cancers, higher concentrations (percentages) are expected in the diagonals of the MZ tables compared to the DZ tables. There is a slight indication of this in the breast-ovarian and breast-cervical combinations but none of the diseases show overwhelming co-aggregation. The bivariate logistic regression analysis was applied to these data. The results from this analysis are presented next.

Table II. Observed frequencies in twin pairs for the traits trouble sleeping and trouble concentrating

	Traits in Twin 2			
Traits in Twin 1	Cancer A =Yes Cancer B =Yes	Cancer A =No Cancer B =Yes	Cancer A =Yes Cancer B = No	Cancer A =No Cancer B =No
Cancer A-Cancer B				
		MZ Twins		
Breast-Cervical				
Yes, Yes	0	0	1	8
No, Yes	0	39	2	205
Yes, No	0	5	35	192
No, No	5	196	194	4569
Breast-Ovarian				
Yes, Yes	0	0	1	0
No, Yes	0	2	1	5
Yes, No	0	3	35	202
No, No	0	43	200	4919
Colon-Rectal				
Yes, Yes	0	0	0	1
No, Yes	0	2	2	38
Yes, No	0	0	8	67
No, No	3	35	65	5230
Colorectal-Endometrial				
Yes, Yes	1	0	1	6
No, Yes	0	4	2	75
Yes, No	0	4	6	86
No, No	2	50	91	9801
		DZ Twins		
Breast-Cervical				
Yes, Yes	0	0	0	11
No, Yes	3	35	10	288
Yes, No	1	8	40	363
No, No	8	328	317	7601
Breast-Ovarian				
Yes, Yes	0	0	0	1
No, Yes	0	2	2	38
Yes, No	0	0	8	67
No, No	3	35	65	5230
Colon-Rectal				
Yes, Yes	0	1	1	5
No, Yes	3	3	2	106
Yes, No	1	7	10	167
No, No	8	96	169	16368
Colorectal-Endometrial				
Yes, Yes	1	0	0	2
No, Yes	0	2	0	86
Yes, No	0	3	9	130
No, No	4	81	114	8581

Results

The table III shows the raw odds ratios for the data on cancer pairs listed in table II. When the MZ odds ratios are significantly different from the DZ odds ratios, there is evidence of additive genetic effect. If the difference between twice the log odds ratio of DZ is significantly larger than the log odds ratio of MZ there is evidence of a shared common environment effect in addition to the additive genetic effect. (See Appendix A for details.) A GEE model that includes the linked trait was fitted to each cancer pair. Results from this analysis are shown in table IV.

Table III. Univariate Odds Ratios for the Individual Cancers and the Linked Trait

		Odds Ratios (95% C. I.)		
Cancer A-Cancer B		Cancer A	Cancer B	Linked trait
Breast-Cervical	MZ	4.38 (2.99, 6.41)	4.37 (3.03, 6.32)	0.60 (.38, 0.95)
	DZ	2.62 (1.87, 3.68)	2.96 (2.08, 4.21)	0.67 (0.48, 0.95)
Breast-Ovarian	MZ	4.38 (2.99, 6.41)	4.95 (1.17, 21.01)	0.55 (0.20, 1.52)
	DZ	2.62 (1.87, 3.68)	2.83 (0.68, 11.73)	1.49 (0.90, 2.44)
Colon-Rectal	MZ	8.71 (4.12, 18.42)	10.61 (4.15, 27.16)	6.15 (3.48, 10.91)
	DZ	5.75 (3.08, 10.77)	5.43 (1.96, 14.98)	4.55 (2.81, 7.37)
Colorectal-Endometrial	MZ	8.92 (4.13, 19.25)	6.89 (1.61, 29.53)	3.26 (1.39, 7.63)
	DZ	5.49 (2.82, 10.70)	3.42 (1.062, 11.02)	1.90 (0.96, 3.76)

The univariate odds ratios seem to indicate additive genetic effects among all cancers. However, in most of the cases the 95% confidence intervals overlap suggesting lack of statistical significance. In terms of linked trait, there seems to be some indication of genetic effect for the Colon-Rectal and the Colorectal-Endometrial pairs.

Table IV. Results from the simultaneous fit of the cancer pairs and linked trait using GEE

	GEE					
Cancer A-Cancer B	Cancer A		Cancer B		Linked Trait	
Effect	χ^2	p-value	χ^2	p-value	χ^2	p-value
Breast-Cervical						
Additive Genetics	3.87	0.049	2.22	0.14	0.17	0.68
Common Environment	1.28	0.26	2.95	0.09	5.15	0.02
Breast-Ovarian						
Additive Genetics	3.87	0.049	0.29	0.59	2.95	0.08
Common Environment	1.28	0.26	0.09	0.77	2.41	0.12
Colon-Rectal						
Additive Genetics	0.69	0.45	0.90	0.34	0.63	0.43
Common Environment	3.21	0.07	0.80	0.37	4.50	0.03
Colorectal-Endometrial						
Additive Genetics	0.87	0.35	0.54	0.46	0.95	0.33
Common Environment	2.40	0.12	0.14	0.71	3.38	0.07

The only cancer that shows statistically significant additive genetic effect is the breast cancer. For the cervical and colon cancers the common environment seems to be marginally significant. For the breast-cervical and colon-rectal combinations the common environment seems to significantly affect the linked trait effects. In other words, there is evidence suggesting that the co-aggregation of breast cancer and cervical cancer is influenced by the twins' shared common environment. Same is true for the colon cancer and the rectal cancer. There is marginal evidence supporting a genetic influence on the co-aggregation of breast cancer and the ovarian cancer.

Discussion.

The research work undertaken in this project led to a bivariate logistic regression analysis for twin data when examining two discrete traits. This multivariate GEE approach, is based on the estimation of odds ratios and provides a method for simultaneous testing of genetic and common environmental hypotheses for single and linked traits. The rationale for use of odds-ratios (and consequently the logistic regression) derives from probabilistic arguments.

The linked trait measure proposed in this article differs from the standard approach used in the literature, which models the conditional probability:

$$\text{P(Trait status of Twin 1} = x \mid \text{Trait status of Twin 2} = y\text{)},$$

where *x, y* assume the value 0 for unaffected and the value 1 for affected. In this specification similar multinomial outcomes are categorized into different categories of affectation. The linked trait definition, on the other hand, looks for disease patterns in a twin and compares them with the disease patterns in the other twin, which makes it easier to interpret.

KEY RESEARCH ACCOMPLISHMENTS:

1. Fully developed methodology for analyzing bivariate outcomes on twins
2. A SAS program written for the applications of the methodology
3. Twin cancer data on breast, ovarian, cervical, Endometrial, colon and rectal cancers were acquired from the Swedish Twin Registry. These data were analyzed using the bivariate logistic regression method proposed.
4. Work on this project has led to two key areas for future research. Although the proposed method has been shown in applications to produce identical results to the classical twin methods such as the Structural Equation Models using tetrachoric correlations, it needs to be formally compared to other methods in the literature using simulations. Second, as mentioned in the conclusions, to perform meaningful bivariate analyses one needs large samples. At this time the only source for such a sample is the Swedish, Finnish, Danish twin registries. While it was possible to acquire data from the Swedish twin registry in the past year, it is tedious to access data from all the three registries. There are no cancer twin registries in the U.S. We are interested in acquiring funds to establish an elderly twin registry. As mentioned in the review section, we have begun some work in this area.

REPORTABLE OUTCOMES

1. A presentation on this topic at the Joint Statistical Meetings, Atlanta, 2001.
2. A poster presented at the ERA of hope conference organized by the DOD in Orlando, 2002.
3. Ph. D. degree (expected completion in December, 2003) for Brandy Rutledge. (This student was partially funded by the project.)
4. An article submitted to the Genetic Epidemiology, (2002).

CONCLUSIONS

Simulation studies have shown (Ramakrishna, 1996), when the prevalence of the disease studied is low or when the sample size is inadequate most univariate twin methods lack power to detect genetic and common environment effects. Further studies are required to determine the sample sizes and prevalence required to achieve reasonable power. One of our students (Brandy Rutledge) is currently pursuing simulation studies to examine power and to compare the method with other classical approaches such as the SEM approach. Our experience suggests, for the twin methods to be applicable, one would require data from large registries that have large samples and also require the prevalence of the diseases to be reasonable. The cell frequencies in most bivariate cancer data are quite small. Therefore, one not only requires large registries, it may also be necessary to combine data from several registries.

The cancer data presented here is from the Swedish Twin registry. Similar twin cancer data have been collected also by the Finnish and Danish twin registries. Including them in the analysis would enormously improve power. The univariate analyses of the cancer twin data that included all the three registries have already been considered and are published (Lichtenstein, et al., 2000). A request has been made to acquire the bivariate data from the Finnish and Danish registries. We hope to analyze these data and publish the results in the future.

In conclusion, an approach to the analysis of discrete bivariate twin data that considers all possible outcomes in the 16 cell multinomial table simultaneously was proposed. The model proposed was formulated in terms of a GEE logistic regression repeated measures framework and applies existing theory and algorithms for estimation and testing. As part of this approach we have defined a linked trait which is of particular interest for testing hypotheses about the familial co-aggregation of disease or illness in twins.

REFERENCES

Ahlbom, A, Lichtenstein, P, Malmstrom, H, Feychting, M, Hemminki, K, Pedersen, NL. 1997. Cancer in Twins: Genetic and nongenetic familial risk factors. Journal of the National Cancer Institute. 89, 287-293.

Bansal, A, Murray, DK, Wu, JT, Stephenson, RA, Middleton RG, Meikle, AW. 2000. Heritability of prostate-specific antigen and relationship with zonal prostate volumes in aging twins. The Journal of Clinical Endocrinology and Metabolism. 85: 1272-1276.

Cerhan, JR, Kushi LH, Olson, Rich SS, Zheng W, Folsom, AR, Sellers TA. 2000. Twinship and Risk of postmenopausal breast cancer. Journal of the National Cancer Institute. 92: 261-264.

Donner A and Klar K. 1996. The statistical analysis of kappa statistics in multiple samples. J Clin Epidemiol 49:1053-8.

Ekbom A, Hsieh CC, Lipworth, L, Adami, HO, Trichopoulos, D. 1997. Intrauterine Environment and Breast Cancer Risk in Women: A population Based Study. Journal of the National Cancer Institute. 89: 71-76.

Falconer DS, MacKay TFC. 1996. Introduction to quantitative genetics. London, England: Benjamin/Cummings Books.

Goldberg J, True WR, Eisen SA, Henderson WG. 1990. A twin study of the effects of the Vietnam War on posttraumatic stress disorder. JAMA 263:1227-32.

Goldberg, J, Henderson, WG, Eisen, SA, True, W, Ramakrishnan, V, Lyons, MJ, Tusang, MT. 1993. A Strategy for Assembling Sample of Adult Twin Pairs in the United States. Statistics in Medicine, 12: 1693 – 1702.

Gronberg, H, Damber, L, Damber, J-E. 1994. Studies of Genetic Factors in Prostate Cancer in a Twin Population. Journal of Urology. 152: 1484-1489.

Hannah MC, Hopper JL, and Mathews JD. 1983. Twin concordance for a binary trait: statistical models illustrated with data on drinking status. Acta Genet Med Gemellol (Roma) 32:127-37.

Hannah MC, Hopper JL, and Mathews JD. 1985. Twin concordance for binary trait II: nested analysis of ever-smoking and ex-smoking traits and unnested analysis of a “committed-smoking○ trait. Amer J Hum Genet 37:153-65.

Harris, EL. 1997. Importance of heritable and nonhertiable variation in cancer susceptibility: evidence from a twin study. The Journal of the National Cancer Institute. 89. 270-276.

Kendler K. 1989. Limitations of the ratio of concordance rates in monozygotic and dizygotic twins. Arch Gen Psychiatry 46:477-8.

Kraemer HC. 1997. What is the right statistical measure of twin concordance (or diagnostic reliability and validity)? Arch Gen Psychiatry 12:1121-4.

Liang K-Y, and Zeger SL. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13-22.

Liang K-Y, Zeger SL and Qaqish B. 1992. Multivariate regression analysis for categorical data. J R Stat Soc, Series B 54:3-40.

Lichtenstein, P, Holm, N. V., Verkasalo P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A and Hemminki, K. 2000. Environmental and Heritable Factors in the Causation of Cancer. Analyses of Cohorts of Twins from Sweden, Denmark and Finland. New England Journal of Medicine, 343, 78 – 85.

Neale MC and Cardon LR. 1992. Methodology for genetic studies of twins and families. Boston: Kluwer.

Page, W, Braun, MM, Partin, AW, Caporaso, N, Walsh, P. 1997. Heredity and Prostate Cancer: A study of World War II veteran twins. The Prostate. 33: 240-245.

Peto, J, Mack, TM. 2000. High constant incidence in twins and other relatives of women with breast cancer. Nature Genetics. 26:411-414.

Ramakrishnan V, Goldberg J, Henderson WG, Eisen SA, True W, Lyons MH, and Tsuang MT. 1992. Elementary methods for the analysis of dichotomous outcomes in unselected sample of twins. Genet Epidemiol 9:273-87.

Ramakrishnan V, Meyer JM, Goldberg J and Henderson WG. 1996. Univariate analysis of dichotomous or ordinal data from twin pairs: a simulation study comparing structural equation modeling and logistic regression. Genet Epidemiol 13:79-90.

Sanderson M, Williams MA, Malone, KE, Stanford, JL, Emanuel, I, White, E, et al. 1996. Perinatal Factors and Risk of Breast Cancer. Epidemiology. 7: 34-37.

Swerdlow, AJ, De Stavola, BL, Swanwick, MA, Maconochie, NES. 1997. Risks of Breast and Testicular Cancers in Young Adult Twins in England and Wales: Evidence on Prenatal and Genetic Aetiology. Lancet. 350: 1723-1728.

Swerdlow, AJ, De Stavola, BL, Swanwick, MA, Mangrani P, Maconochie NES. 1999. Risk Factors for Testicular Cancer: A Case-control Study in Twins. Cancer Causes and Control. 80: 1098-1102.

Terry, P, Baron, JA, Weiderpass, E, Yuen J, Lichtenstein, P and Nyren, O. 1999. Lifestyle and Endometrial Cancer Risk: A Cohort Study from the Swedish Twin Registry. Internation Journal of Cancer: 82, 38-42.

Verkasalo, PK, Kaprio, J, Pukkala, E, Koskenvuo, M. 1999. Breast Cancer Risk in Monozygotic and Dizygotic Female Twins: A 20-year Population-based Cohort Study in Finland from 1976 to 1995. Cancer Epidemiology, Biomarkers and Prevention. 8: 271-274.

Verkasalo, PK, Kaprio, J, Pukkala, E, Koskenvuo, M. 1999. Breast Cancer Risk in Monozygotic and Dyzygotic Female Twins: A 20-Year Population-based Cohort Study in Finland from 1976-1995. Cancer, Epidemiology, Biomarkers and Prevension. 80: 1098 – 1102.

Verkasalo, PK, Kaprio, J, Pukkala, E, Koskenvuo, M. 1999. Genetic Predisposition, Environment and Cancer Incidence: A Nationwide Twin Study in Finland, 1976-1995. International Journal of Cancer. 83: 743-749.

Zeger SL and Liang K-Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42:121-30.

Zeger SL and Liang K-Y. 1992. An overview of methods for the analysis of longitudinal data. Stat Med 11:1825-39.

Appendix A

An Application of Generalized Estimating Equations Method for Twin Data

V. Ramakrishnan[1], William Henderson[2] Jack Goldberg[3] ,Tammy Massie[1]

[1] Virginia Commonwealth University, Department of Biostatistics, Richmond, VA (V. R., T. M)
[2] VA Cooperative Studies Program Coordinating Center, VA Hospital, Hines, IL (W. G. H.)
[3]Seattle VA ERIC/Vietnam Era Twin Registry and the University of Washington, Department of Epidemiology, Seattle, WA (J. G.)

Author for Correspondence:
Tammy Massie
Department of Biostatistics
Virginia Commonwealth University
1101 E Marshall St.
Richmond, VA, 23298-0032
E-mail: tjparlim@hsc.vcu.edu
Tel: 804 827 2040
Fax 804 828-8900

Abstract

This paper presents a new method for the simultaneous genetic analysis of two or more discrete traits by applying a Generalized Estimating Equations (GEE) approach to fitting logistic regression model for twin data. Co-aggregation between the two traits is estimated using an odds ratio and justifications for this measure are provided. This *linked trait* is defined for two discrete traits based upon explicit patterns of trait concordance and discordance within twin pairs; this linked trait is assessed for the influence of additive genetic and/or common environmental effects. Data are summarized in the form of 2 x 2 tables (for monozygotic and dizygotic twins) by combining appropriate cells from the 16-cell multinomial distribution to define the individual and linked trait. Hypothesis tests comparing monozygotic and dizygotic twins are performed using repeated measures logistic regression via the GEE approach. The model specification is highly flexible, accounts for the correlated structure of the parameter estimates and does not require multivariate normality assumptions for the underlying distribution. The approach is illustrated using two example data sets from the Vietnam Era Twin (VET) Registry.

1. Introduction

The statistical methodology for determining how genes influence the occurrence of disease is a topic of interest among both quantitative geneticists and genetic epidemiologists. Data from twin studies can be used to examine how complex genetic mechanisms and non-genetic factors influence disease occurrence. The classical twin study has long been used to estimate genetic influence on a single trait. This research design compares within-pair twin similarity for a trait in monozygotic (MZ) and dizygotic (DZ) pairs; from these similarity measures estimates of the influence of heredity and common environment are developed. The methods for the analysis of classical twin studies of continuously distributed traits (height, weight, etc) are well described and use the intra-class correlation as the principal indicator of twin similarity [Falconer and MacKay, 1996]. However, for discrete data there is no universally accepted measure of twin similarity [Hannah et al., 1983; 1985; Kendler, 1989; Neale and Cardon, 1992; Donner, 1996; Kraemer, 1997].

The odds ratio derived from logistic regression has been proposed as a measure of twin similarity in classical twin studies [Ramakrishnan et al., 1992; 1996]. The odds ratio when used with purely categorical data does not require the assumption of an underlying bivariate normal distribution. After an initial assessment of the genetic basis of a single trait it is often of interest to examine the co-aggregation of more than one trait or disease within twin pairs. For example, one might want to explore the co-occurrence of two types of cancer (breast and ovarian) or two types of substance abuse (alcohol and illicit drugs) in pairs of twins. Twins are especially useful for these types of analyses since it is possible to examine if two traits are influenced by shared additive genetic effects. However, to address the co-aggregation of two traits the univariate methods of twin analysis have to be extended to the bivariate case.

In this paper, a new method for the simultaneous genetic analysis of two or more discrete traits is proposed using a GEE logistic regression model [Liang et al., 1992; Liang and Zeger, 1986; Zeger and Liang, 1986; 1992] that is adapted for twin data. The remainder of this paper consists of four sections. First, the theoretical rationale for the use of the odds ratio as the measure of twin similarity is presented. Second, a series of `univariate' logistic regression models that permits hypothesis testing for genetic and common environmental effects with a single discrete trait are reviewed. Third, the univariate logistic model is extended to the `bivariate' case to examine if two discrete traits share a common genetic or familial vulnerability. Fourth, the approach is illustrated using data from the Vietnam Era Twin (VET) Registry [Henderson et al., 1990].

2. Twin Data Methods.

2.1. Rationale for a Genetic Analysis of Twin Data Using Odds Ratios

When a binary discrete trait is considered, twins can be classified into three possible groups: concordant for the trait, discordant for the trait and concordant for the absence of the trait. If individuals in a population are paired at random then the expected probabilities of the three groups, say p_1, p_2 and p_3, would conform to a probably law (that is somewhat similar to the Hardy-Weinberg Law for random matings). That is, if p denotes the probability that an individual carries the trait, then under random-pairing, $p_1 = p^2, p_2 = 2pq$ and $p_3 = q^2$. This yields $4p_1p_3 = p_2^2$ or equivalently $\psi = 4p_1p_3 / p_2^2 = 1$, where ψ is the odds ratio under the condition of exchangeability (namely, the probabilities of the discordant cells are equal).

The expected patterns [p^2, $2pq$ and q^2] for concordant affected, discordant, and concordant unaffected twins are formulated without regard for the zygosity of the twins. However, it is also possible to derive separate estimates of these patterns and accompanying odds ratios in MZ and DZ pairs. If these two odds ratios are significantly different it would indicate evidence of genetic effect. In addition, in the classical twin analyses, the fact that on average DZ twins share half as much of their genetic material as MZ twins is used to separate the *additive* genetic effect from *shared common environment* effect. Since the effects are multiplicative in terms of the odds ratios, when the MZ log odds ratio is twice the DZ log odds ratio the evidence is deemed purely due to genetics. If the DZ log odds ratio is larger than twice the MZ log odds ratio, the additional effect is attributed to the shared common environment (in the presence of the genetic effect, which is called the additive genetic effect). Hypothesis tests could be developed for both the genetic effect and the shared common environment effect. These arguments are analogous to those applied to the models for the analyses of continuous data in twins [Falconer, 1965; Falconer and MacKay, 1996; Defries et al., 1987].

2.2. Univariate Genetic Analysis of Twin Odds Ratios Using Logistic Regression

Let *Y* and *X* denote the binary traits for the members of a twin pair (arbitrarily referred to as twin 1 and twin 2), respectively. Let Z denote the zygosity of the twin pair, where the MZ pairs are coded as 1 and the DZ pairs are coded as 0.5. Let *XZ* denote the product of *X* and *Z*. Then the model for the logit of the conditional probability of *Y* given the outcome for *X* and *Z* is written

$$Ln\left[\frac{P(Y=1|X,Z)}{1-P(Y=1|X,Z)}\right]=\beta_0+\beta_1 X+\beta_2 Z+\beta_3 XZ. \tag{1}$$

The choice of (1, 0.5) coding for *Z* is to facilitate certain specific hypotheses regarding MZ and DZ log-odds ratios. Consider the model for MZ twins, which is obtained by substituting Z = 1 in (1). The model is

$$Ln\left[\frac{P(Y=1|X,Z=1)}{1-P(Y=1|X,Z=1)}\right]=(\beta_0+\beta_2)+(\beta_1+\beta_3)X, \tag{2}$$

where $\beta_{MZ}=\beta_1+\beta_3$ represents the log-odds ratio that estimates the association between the MZ twins. Similarly consider the model for the DZ twin by substituting Z = 0.5 in (1), which yields

$$Ln\left[\frac{P(Y=1|X,Z=0.5)}{1-P(Y=1|X,Z=0.5)}\right]=(\beta_0+0.5\beta_2)+(\beta_{12}+0.5\beta_3)X, \tag{3}$$

where $\beta_{DZ}=\beta_1+0.5\beta_3$ represents the log-odds ratio that estimates the association between the DZ twins.

From the regression equations defined in 2 and 3 above the following hypotheses are of interest:

i) The trait observed is due to an additive genetic effect. This could be tested using the null and alternative hypotheses as follows.

$$H_0 : \beta_3 = 0$$
$$H_a : \beta_3 > 0$$

That is, the null hypothesis states that the MZ log-odds ratio is equal to the DZ log-odds ratio of the DZ twins. If there is no evidence to support an additive genetic effect the hypothesis of familial clustering could be tested next. This is done by refitting a model with only the co-twin variable (*X*) in the model and testing whether the corresponding coefficient is zero. (That is, $\beta_1 = 0$.)

ii) The trait observed is due to both and additive genetic effects and common environment effects. This could be tested by specifying the following null and alternative hypotheses:

$$H_0 : \beta_1 \leq 0$$
$$H_a : \beta_1 > 0$$

Under model (1) the null hypothesis states that the MZ log-odds ratio is equal to two times the DZ log-odds ratio. If a trait were purely due to additive genetic effects, then the ratio of the MZ to DZ log-odds ratios would approximate the 2 to 1 ratio of the MZ to DZ ratio of genes, respectively.

Others [Gao et al., 1997; Betensky et al., 2001] have pointed out that the designation of twin 1 and twin 2 (or index twin vs. co-twin) is problematic for the odds ratio estimation procedure. They note that the discordant cell probabilities are formally exchangeable and that in small samples an arbitrary designation of twin 1 and twin 2 could lead to inconsistent results. One solution to this problem is to use a repeated measure logistic regression model [Connolly and Liang, 1988; Betensky et al., 2001] fitted using the GEE approach. An equivalent and somewhat simpler solution is to distribute exactly half of the total number of trait discordant pairs into the two discordant cells and than apply a standard logistic regression model. Simulation studies comparing the logistic regression method with other methods are also available in the literature (Ramakrishnan, et al, 1996, Gao et al., 1997). (Most statistical packages that perform logistic regression analysis allow the data to be entered in tabular format.)

2.3. Bivariate Genetic Analysis of Twin Odds Ratios Using Logistic Regression

2.3.1 Rationale for the bivariate genetic analysis

The rationale for the bivariate analysis flows directly from the Hardy-Weinberg Law as it is extended to two traits. Suppose A and B indicate that twin 1 is affected with both traits and $\bar{A}$ and $\bar{B}$ denotes that twin 1 is unaffected by both traits. Let π_A and π_B denote the separate probabilities that traits A and B are present in an individual; π_{AB} represents the probability that *both* traits are present. Under the null-hypothesis of independence the relationship $\pi_{AB}=\pi_A \pi_B$ will hold. Similarly, $\pi_{A\bar{B}}=\pi_A(1-\pi_B)$, $\pi_{\bar{A}B}=(1-\pi_A)\pi_B$ and $\pi_{\bar{A}\bar{B}}=(1-\pi_A)(1-\pi_B)$ will hold.

Now consider the data layout for two traits in twins as defined in table I. Appropriate combinations of the observed frequencies of the 16 cells in table I will produce the corresponding 2 x 2 tables that can be used to analyze traits A and B. The cell combinations with the corresponding cell probabilities under the null hypothesis of independence for traits A and B are shown in Table II.

Insert Table I here

Insert Table II here

To study the co-aggregation of the two traits, a quantity called the *linked trait* is defined. Evidence of a linked trait is observed when the to twins responses agree for both trait A, and trait B. This is termed concordant-affected for linked trait. Similarly, if the twin responses disagree for both traits, there is evidence against linked trait. This is termed concordant-unaffected for linked trait absent. Other events are termed discordant for linked trait.

Comparisons of cross-twin cross-trait affectation statuses for the two traits *A* and *B* can be represented in the form of a 2 x 2 contingency table similar to the tables for the traits A and B (Table IIc). The elements (m_{11}, m_{22}, m_{33}, m_{44}) represent the frequencies of events that are concordant-affected for the linked trait. Similarly, the elements (m_{14}, m_{23}, m_{32}, m_{41}) represent the frequencies of events that are concordant-unaffected for the linked trait. The remaining cells correspond to discordance for the linked trait. A measure of the linked trait (LT), say ψ_{LT}, is defined by the following statistic:

$$\psi_{LTE} = \frac{P(concordant\ affected\ for\ LT)P(concordant\ unaffected\ for\ LT)}{P(discordant\ for\ LT)} \qquad (4)$$

Consider now the probabilities in the expression for ψ_{LT} given in the above equation (4). Denoting the probability *P* [concordant-affected LT] as p_{11},

$p_{11} = P$ [(Twin 1 has *A* and *B* *and* Twin 2 has *A* and *B*) *or*

(Twin 1 has *A* but not *B* *and* Twin 2 has *B* but not *A*) *or*

(Twin 1 has *B* but not *A* *and* Twin 2 has *A* but not *B*) *or*

(Twin 1 does not have *A* or *B* *and* Twin 2 does not have *A* or *B*)].

Under the null hypothesis that no linked trait is present, the twins will resemble randomly paired individuals and the occurrence of traits will be independent. The independence between twins yields, $p_{11} = \pi_{AB}\pi_{AB} + \pi_{A\bar{B}}\pi_{\bar{A}B} + \pi_{\bar{A}B}\pi_{A\bar{B}} + \pi_{\bar{A}\bar{B}}\pi_{\bar{A}\bar{B}}$. Applying the independence of traits (as argued earlier) will yield

$$p_{11} = (\pi_A^2 + (1-\pi_A^2))(\pi_B^2 + (1-\pi_B^2)).$$

Similarly,

$$P\ [\text{concordant-unaffected LT}] = p_{00} = 4\pi_A\pi_B(1-\pi_A)(1-\pi_B),$$

$$P\ [\text{discordant for LT when the twins agree on trait A}] = p_{10} = 2\pi_A(1-\pi_A)(\pi_B^2 + (1-\pi_B)^2)$$

and

$$P\text{ [discordant for LT when the twins agree on trait B]} = p_{01} = 2\pi_B(1-\pi_B)(\pi_A^2 + (1-\pi_A)^2)$$

Notice that the equation $p_{11}p_{00} = p_{10}p_{01}$ holds. Consequently, if there is no linked trait then the odds ratio,

$\Psi_{ST} = \dfrac{p_{11}p_{22}}{p_{12}p_{21}} = 1$. A large value of Ψ_{ST} would indicate the lack of independence of the traits within the twin pairs suggesting the presence of a linked trait. Similar to the formulation for single traits, additional analysis seeks to compare the MZ and DZ odds ratios to test for the influence of additive genetics and common environment.

2.3.2. Modeling the linked trait: estimation of parameters

The estimate of the linked trait odds ratio can be obtained by maximizing the multinomial likelihood and substituting the various estimates of the probabilities applying the invariance of maximum likelihood estimators (mle's). However the mle's of the odds ratios representing the two traits and the linked trait are correlated. Because of this correlation it is necessary to obtain these estimates simultaneously along with the appropriate variance-covariance matrix.

Simultaneous parameter estimation is accomplished using a repeated measures analysis with the GEE method. To implement the GEE method two new indicator variables are defined:

L_A = 1 if the two twins agree for trait A

= 0 otherwise

L_B = 1 if the two twins agree for trait B

= 0 otherwise

The various combinations of L_A and L_B are described in Table III.

Insert Table III here

With these two new indicator variables the odds ratio for the linked trait in equation (4) is nothing more than the odds ratio between L$_A$ and L$_B$. The odds ratio will be larger than 1 when there is evidence in support of a linked trait; further, if the linked trait is due to additive genetics then the MZ odds ratio will be larger than the DZ odds ratio.

Using logistic regression and defining the tables listed above for MZ and DZ twins separately, one can obtain simultaneous estimates of the various parameters of interest by fitting the three logistic regression models.

$$Ln\left[\frac{P(Y_A = 1|X_A, Z = 1)}{1 - P(Y_A = 1|X_A, Z = 1)}\right] = \beta_0^A + \beta_1^A X_A + \beta_2^A Z + \beta_3^A X_A,$$

$$Ln\left[\frac{P(Y_B = 1|X_B, Z = 1)}{1 - P(Y_B = 1|X_B, Z = 1)}\right] = \beta_0^B + \beta_1^B X_B + \beta_2^B Z + \beta_3^B X_B, \tag{5}$$

$$Ln\left[\frac{P(L_A = 1 | L_B, Z = 1)}{1 - P(L_A = 1 | L_B, Z = 1)}\right] = \beta_0^L + \beta_1^L L_B + \beta_2^L Z + \beta_3^L L_B.$$

Substituting $Z = 1$ for the MZ twins equation (8) yields the log odds ratio $\beta^i_{MZ} = \beta_1^i + \beta_3^i$ for the MZ twins. Similarly, substituting $Z = 0.5$ for the DZ twins equation (5) yields the log-odds ratio $\beta^i_{DZ} = \beta_1^i + 0.5\beta_3^i$ for the DZ twins. Here, $i = A, B$ or L, for traits A, B and the linked trait, respectively.

To avoid the index-twin, co-twin labeling the models in (5) additional three models with the twin variables interchanged should be considered. Several software packages (SAS, STATA and MIXOR [Hedeker, 1993]) are available for fitting repeated measures models that would be appropriate for implementing the estimation procedures for bivariate twin analysis. We propose using PROC GENMOD with the REPEATED option in SAS, which uses the GEE method. To do this, the above three models have to be reformulated into a single logistic regression model. The data structure and SAS syntax for performing this analysis is described in the appendix.

3. Data from the Vietnam Era Twin (VET) Registry

The data used to illustrate the methodology are from the Vietnam Era Twin (VET) Registry [Henderson et al., 1990]. The Registry was assembled from Department of Defense (DoD) computer records and consists of male-male twin pairs who both served in the military during the Vietnam era (between 1965-1975) and were born between 1939-1957. The median year of birth for twins on the Registry is 1949 and over 90% of the twins are white. In total ~7,500 twin pairs were initially identified from DoD files.

The first data set used to illustrate the bivariate analysis is derived from a mail and telephone survey of the twins conducted in 1987 and contains data on ~4500 complete twin pairs (55% are MZ and 45% are DZ). Data collected included several symptoms of posttraumatic stress disorder (PTSD) [Goldberg et al., 1990]. Two symptoms, ‘trouble sleeping’ and ‘trouble concentrating’, are used to illustrate our approach. While these measures were originally collected as five level ordinal variables rating the frequency of PTSD symptoms we have dichotomized the responses (present = 0, 1 or 2 and absent = 3 or 4). The second data set includes two measures of smoking and alcohol drinking derived from a 1990 mail and telephone survey to identify risk factors for cardiovascular disease [Fabsitz et al., 1997]: lifetime cigarette smoking is defined as having smoked at least 100 cigarettes and lifetime alcohol drinking is defined as having had at least 20 drinks. Data on smoking and drinking variables were available on 3025 pairs.

3.1. Analysis of PTSD symptom data

Table IV shows the frequencies for trouble sleeping and trouble concentrating in both twin pairs. The clustering along the main diagonals is an indication of the familial co-aggregation for these two traits. The largest frequencies occur in the concordant cells. This indicates the presence of twin effects for each characteristic as well as evidence for a linked trait.

Insert Table IV here

Table V displays the estimated MZ and DZ odds ratios and standard errors from the GEE repeated measures model. The MZ odds ratios are consistently larger than the DZ odds ratios indicating a genetic trend.

Insert Table V here

Hypothesis tests based on likelihood ratio Chi-squared tests with 1 df are presented in table VI for additive genetics and common environment. The test statistics suggest that trouble sleeping and trouble concentrating are in part under genetic control; there is little evidence for a common environmental influence on these traits. The tests for genetic influence on the linked trait are insignificant suggesting that there may not be any co-aggregations influencing the two traits simultaneously.

Insert Table VI here

3.3. Cigarette smoking and alcohol drinking

The distribution of the cigarette smoking and alcohol drinking across the twin pairs is presented in table VII. The largest frequencies occur in the cells concordant for the three effects. However, the distribution of cell frequencies in the two tables seems similar indicating lack of evidence for co-aggregation of the two traits. Also the off-diagonal cells are sparse and in the DZ table there is a zero cell. This often leads to problems in fitting GEE models with an unspecified working correlation matrix. An 'exchangeable' matrix was chosen in this case.

Insert Table VII here

Table VIII shows the estimated twin odds ratios from the GEE repeated measures model. MZ twin odds ratios are substantially greater than the comparable DZ twin odds ratio for both traits suggesting the influence of genetic factors.

Insert Table VIII here

Formal testing for genetic and common environmental effects for cigarette smoking, alcohol drinking and the linked trait are presented in Table IX. There is strong evidence of genetic effect for both cigarette smoking and alcohol drinking. There is also a strong evidence of shared common environment effect for drinking however this effect is not statistically significant for cigarette smoking. The linked trait effect is not statistically significant leading to the conclusion that there is no evidence of co-aggregation between the two traits.

Insert Table IX here

4. Discussion

This paper presents a bivariate logistic regression analysis for twin data when examining two discrete traits. This multivariate GEE approach, is based on the estimation of odds ratios and provides a method for simultaneous testing of genetic and common environmental hypotheses for single and linked traits. The rationale for use of odds-ratios (and consequently the logistic regression) derives from probabilistic arguments.

In a similar vein Betensky et al., [2001], have recently proposed a GEE approach for the analysis of discrete bivariate twin data. However, their model specification differs from the one presented by us in this article. They model the conditional probability:

$$\text{P(Trait status of Twin 1} = x \mid \text{Trait status of Twin 2} = y\text{)},$$

where *x, y* assume the value 0 for unaffected and the value 1 for affected. In this specification similar multinomial outcomes are categorized into different categories of affectation. For instance, consider the concordant-affected status for both traits (i.e., this corresponds to P (Trait status of Twin 1 = 1 and Trait status of Twin 2 = 1)). The outcomes are

$$\{ABAB, AB\,\overline{A}\,B, A\,\overline{B}\,AB, A\,\overline{B}\;\;\overline{A}\,B\}.$$

Now consider concordant-unaffected status for both traits (i.e., P (Trait status of Twin 1 = 0 and Trait status of Twin 2 = 0)). The outcomes are

$$\{\overline{A}BA\overline{B}, \overline{A}B\overline{A}\,\overline{B}, \overline{A}\,\overline{B}A\overline{B}, \overline{A}\,\overline{B}\,\overline{A}\,\overline{B}\}.$$

Since the labeling of the twins is arbitrary, the outcomes $A\overline{B}\,\overline{A}B$ and $\overline{A}BA\overline{B}$ essentially convey the same information. However, these two outcomess appear in two different probabilities in the Betensky et al. [2001] approach. Similarly, the outcomes $\overline{A}B\overline{A}\,\overline{B}$ and $\overline{A}\,\overline{B}\,\overline{A}B$ are equivalent but the former appears in the concordant-unaffected cell and the latter appears in the discordant cell. This specification makes it difficult to interpret the corresponding effect as a linked trait effect. The linked trait defined here is one way to simultaneously compare the occurrence of the traits in the two twins. Other definitions might be considered. For instance, one may compare the occurrence of trait A in twin 1 with the occurrence of trait B in twin 2 (Bhadra, 1999). Once the 2 x 2 tables are constructed using this definition the methodology would be identical. When the prevalences of the two traits are similar most of these methods would lead to similar results.

An intra-class correlation method suggested as an alternative to the univariate logistic regression method [Gao et al., 1996] could also be extended to the bivariate case. Currently such an extension is not available. Moreover, the method does not easily allow for testing the common environment effect in the univariate case. It would be even more difficult to formulate the appropriate tests for the bivariate case.

Several extensions to the method proposed here are possible. Using multinomial theory it would be straightforward to extend the methodology beyond the bivariate situation. The rationale would be motivated by deriving probabilities and twin odds ratios using a probability law (similar to the Hardy-Weinberg Law for random mating). For instance, the trivariate case would require two bivariate linked traits and one trivariate linked trait. Even though the interpretation of the higher order effects would be complicated, the methodology easily lends itself to such extensions. Theoretically, the method could be extended to ordinal outcomes provided the proportional odds assumption [McCullagh, 1980] holds. However, adapting standard computer software packages (like SAS and STATA) for the genetic analysis of twin data with ordinal outcomes using the GEE approach is not currently available. Alternative procedures for the analysis of ordinal outcomes in twins, such as mixed models using the MIXOR software package [Hedeker et al., 1993], the Alternating Logistic Regression method should be investigated. Inclusion of covariates is a simple extension in the bivariate GEE twin analysis. Covariates for the pair level, such as age and race, are especially appealing and could be added to the model as adjustment factors. Further, by adding interaction terms for age and race it would be possible to assess if there are differential genetic influences on each trait and the linked trait.

Further work is also needed to assess the power of the GEE approach for the detection of the linked trait, particularly with respect to common environmental influences. Previously, we have shown that for the univariate analysis [Ramakrishnan et al., 1996], the logistic regression method has only modest power to detect a common environmental effect when it is present. As the number of traits increases this problem could get worse.

In conclusion, we have proposed an approach to the analysis of discrete bivariate twin data that considers all possible outcomes in the 16 cell multinomial table simultaneously. The model we propose is formulated in terms of a GEE logistic regression repeated measures framework and applies existing theory and algorithms for estimation and testing. As part of this approach we have defined a linked trait which is of particular interest for testing hypotheses about the familial co-aggregation of disease or illness in twins.

ACKNOWLEDGEMENTS

Department of Veterans Affairs Health Services and Development Service: Chief Research and Development Officer, John R. Feussner, M.D.; Cooperative Studies Program, Assistant Director: Steven M. Berkowitz, Ph.D; Cooperative Studies Program, Seattle VA Epidemiologic Research and Information Center, Vietnam Era Twin Registry: Director, Edward Boyko, MD, MPH; Registry Coordinator, Birute Curran; Vietnam Era Twin Registry Advisory Committee: Theodore Colton, Sc.D., Ralph Paffenbarger, M.D., Walter Nance, M.D., Myrna Weissman, Ph.D., Emil Cacarro, M.D. Dr. Irving Gottesman made important contributions to the development of the Vietnam Era Twin Registry. Most importantly, the authors wish to acknowledge and thank the members of the Vietnam Era Twin Registry for their participation and cooperation. They willingly provided sensitive information and considerable time in responding to the surveys. Without their contribution this research project would not have been possible. **This research was funded by the Department of Defense.**

REFERENCES

Betensky RA, Hudson JI, Jones CA, Hu F, Wang B, Chen C, Xu X. 2001. A computationally simple test of homogeneity of odds ratios for twin data. Genet Epidemiol 20:228-38.

Bhadra, P. (1999), Genetic analysis of bivariate dichotomous twin-data using logistic regression models. Doctoral Dissertation, University of Chicago, Chicago, Illinois.

Connolly MA, Liang K-Y. 1988. Conditional logistic regression models for correlated binary data. Biometrika 75:501-6.

DeFries JC and Fulker DW. 1985. Multiple regression analysis of twin data. Behav Genet 15:467-73.

Donner A, Klar N, and Eliasziw M. 1995. Statistical methodology for estimating twin similarity with respect to a dichotomous trait. Genet Epidemiol 12:267-77.

Donner A and Klar K. 1996. The statistical analysis of kappa statistics in multiple samples. J Clin Epidemiol 49:1053-8.

Fabsitz RR, Sholinsky P, Goldberg J. 1997. Correlates of sleep problems among men: the Vietnam Era Twin Registry. J Sleep Res 6:50-6.

Falconer DS. 1965. The inheritance of liability to certain diseases estimated from the incidence among relatives. Ann Hum Genet 29:51-76.

Falconer DS, MacKay TFC. 1996. Introduction to quantitative genetics. London, England: Benjamin/Cummings Books.

Gao XJ, Klar N, and Donner A. 1997. Comparison of methods for analyzing binary data arising from two-sample twin studies. Genet Epidemiol 14:349-63.

Glonek GFV and McCullagh P 1995. Multivariate logistic models. J R Stat Soc, Series B 57:533-46.

Goldberg J, True WR, Eisen SA, Henderson WG. 1990. A twin study of the effects of the Vietnam War on posttraumatic stress disorder. JAMA 263:1227-32

Hannah MC, Hopper JL, and Mathews JD. 1983. Twin concordance for a binary trait: statistical models illustrated with data on drinking status. Acta Genet Med Gemellol (Roma) 32:127-37.

Hannah MC, Hopper JL, and Mathews JD. 1985. Twin concordance for binary trait II: nested analysis of ever-smoking and ex-smoking traits and unnested analysis of a "committed-smoking○ trait. Amer J Hum Genet 37:153-65.

Hedeker D and Gibbons RD. 1994. A random-effects ordinal regression model for multilevel analysis. Biometrics 50:933-44.

Henderson WG, Eisen S, Goldberg J, True WR, Barnes JE, Vitek ME. 1990. The Vietnam Era Twin Registry: a resource for medical research. Public Health Rep 105:368-73

Hrubec Z and Robinette CD. 1984. The study of human twins in medical research. N Engl J Med 310:435-41.

Kendler K. 1989. Limitations of the ratio of concordance rates in monozygotic and dizygotic twins. Arch Gen Psychiatry 46:477-8.

Khoury MJ, Beaty TH and Cohen BH. 1993. Fundamentals of genetic epidemiology, New York: Oxford.

Kraemer HC. 1997. What is the right statistical measure of twin concordance (or diagnostic reliability and validity)? Arch Gen Psychiatry 12:1121-4.

Liang K-Y, and Zeger SL. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13-22.

Liang K-Y, Zeger SL and Qaqish B. 1992. Multivariate regression analysis for categorical data. J R Stat Soc, Series B 54:3-40.

Massie, TJ. 2002. Testing Genetic Hypothesis on Bivariate Dichotomous Twin Data Using Repeated Measures Logistic Regression . Doctoral Dissertation, MCV-VCU, Richmond, Va.

McCullagh P. 1980. Regression models for ordinal data. J R Stat Soc, Series B 42:109-42.

McCullagh P. 1989. Models for discrete multivariate responses. Bulletin of the International Statistical Institute 53:407-18.

Neale MC and Cardon LR. 1992. Methodology for genetic studies of twins and families. Boston: Kluwer.

Olson, JM. 1992. Testing the Hardy-Weinberg law across strata. Ann Hum Genet 57:291-5.

Olson JM, Witte JS and Elston RC. 1996. Association within twin pairs for a dichotomous trait. Genet Epidemiol 13:489-99.

Ramakrishnan V, Goldberg J, Henderson WG, Eisen SA, True W, Lyons MH, and Tsuang MT. 1992. Elementary methods for the analysis of dichotomous outcomes in unselected sample of twins. Genet Epidemiol 9:273-87.

Ramakrishnan V, Meyer JM, Goldberg J and Henderson WG. 1996. Univariate analysis of dichotomous or ordinal data from twin pairs: a simulation study comparing structural equation modeling and logistic regression. Genet Epidemiol 13:79-90.

SAS Institute Inc. 1987. SAS Users Guide: Statistics,Version 6.12 edition. Cary, North Carolina: SAS Institute Inc.

Sham PC, Walters EE, Neale MC, Heath AC, MacLean CJ and Kendler KS. 1994. Logistic regression analysis of twin data: estimation of parameters of the multifactorial liability threshold model. Behav Genet 24:229-38.

True WR, Rice J, Eisen SA, Heath AC, Goldberg J, Lyons MJ and Nowak J. 1993. A twin study of genetic and environmental contributions to liability for posttraumatic stress symptoms. Arch Gen Psychiatry 50:257-64.

Zeger SL and Liang K-Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42:121-30.

Zeger SL and Liang K-Y. 1992. An overview of methods for the analysis of longitudinal data. Stat Med 11:1825-39.

APPENDIX

GEE Computer Program Specifications for the Bivariate Twin Logistic Regression Model

The SAS PROC GENMOD procedure was used to estimate and test parameters in the bivariate logistic regression model. The data structure was as follows.

Pair ID	Twin 1	Twin 2	Zygosity	Indicators	
(ID)	(Y)	(X)	(Z)	D_1	D_2
1	Y_A	X_A	1	1	0
1	Y_B	X_B	1	0	1
1	L_A	L_B	1	0	0
1	X_A	Y_A	1	1	0
1	X_B	Y_B	1	0	1
1	L_B	L_A	1	0	0
2	Y_A	X_A	1	1	0
2	Y_B	X_B	1	0	1
2	L_A	L_B	1	0	0
2	X_A	Y_A	1	1	0
2	X_B	Y_B	1	0	1

2	L_B	L_A	1	0	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n_1	Y_A	X_A	1	1	0
n_1	Y_B	X_B	1	0	1
n_1	L_A	L_B	1	0	0
n_1	Y_A	X_A	1	1	0
n_1	Y_B	X_B	1	0	1
n_1	L_B	L_A	1	0	0
n_1+1	Y_A	X_A	.5	1	0
n_1+1	Y_B	X_B	.5	0	1
n_1+1	L_A	L_B	.5	0	0
n_1+1	Y_A	X_A	.5	1	0
n_1+1	Y_B	X_B	.5	0	1
n_1+1	L_B	L_A	.5	0	0
n_1+2	Y_A	X_A	.5	1	0
n_1+2	Y_B	X_B	.5	0	1
n_1+2	L_A	L_B	.5	0	0

n_1+2	Y_A	X_A	.5	1	0
n_1+2	Y_B	X_B	.5	0	1
n_1+2	L_B	L_A	.5	0	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n_1+n_2	Y_A	X_A	.5	1	0
n_1+n_2	Y_B	X_B	.5	0	1
n_1+n_2	L_A	L_B	.5	0	0
n_1+n_2	Y_A	X_A	.5	1	0
n_1+n_2	Y_B	X_B	.5	0	1
n_1+n_2	L_B	L_A	.5	0	0

The variables Y and X are coded as binary (1 for presence and 0 for absence). The subscripts A, B and L represent the two traits and the linked trait, respectively. The coding of zygosity as 1 and 0.5 is to facilitate testing for the common environment in the presence of additive genetic effects. The indicator variables D_1 and D_2 specify the outcomes A, B, or L in the logistic regression model. For example, the pattern $D_1 = 1$ and $D_2 = 0$ corresponds to trait A, $D_1 = 0$ and $D_2 = 1$ corresponds to trait B, and $D_1 = 0$ and $D_2 = 0$ corresponds to the linked trait. The logistic regression equation for fitting the newly formatted data is given by:

$$Ln\left[\frac{P(Y=1|X,Z,D_1,D_2)}{1-P(Y=1|X,Z,D_1,D_2)}\right] = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 D_1 + \beta_4 D_2 + \beta_5 XZ + \beta_6 XD_1 + \beta_7 XD_2 + \beta_8 ZD_1 + \beta_9 ZD_2 + \beta_{10} XZD_1 + \beta_{11} XZD_2. \qquad (9)$$

By substituting various values for D_1 and D_2 there is a formal equivalence between the equations (9) and (5). For example, $\beta_0^A = \beta_0 + \beta_3$, $\beta_1^A = \beta_1 + \beta_6$, $\beta_2^A = \beta_2 + \beta_8$, $\beta_3^A = \beta_5 + \beta_{10}$ and so on for B and L. The tests for additive genetic effects and common environment effects for A, B and L can be performed by score tests on linear combinations of the parameters. These tests would be χ_1^2 tests. With no prior knowledge of the correlation among the outcomes, an 'unstructured' working correlation matrix could be used in the GEE model. An unstructured correlation matrix assumes that the correlations between various pairs of outcomes have different values. In some situation the 'exchangeable' matrix could also be used. The exchangeable matrix assigns the same correlation for all off-diagonal entries in the working correlation matrix. Model selection based on either the unstructured or the exchangeable correlation matrix leads to consistent estimators. However, one closer to the true structure might lead to higher efficiency for the estimates of the regression parameters.

SAS syntax for performing the bivariate genetic analysis:

```
PROC GENMOD DESCENDING;
                         CLASS IDNUM;
/*** CLASS statement identifies the subjects as the cluster in which the measurements are
repeated ***/;
MODEL INDEXTWIN= COTWIN  Z  D1 D2 ZCO D1CO D2CO D1Z D2Z D1ZCO D2ZCO /
DIST=B LINK=LOGIT;
/*** The MODEL statement specifies the indextwin as the logit of the response variable Y ***/;
/*** ZCO, D1CO, D2CO, D1Z, D2Z, D1ZCO and D2ZCO are interaction terms, where
COTWIN (CO) represents the X term in equation (9) ***/;
/*** DIST = B specifies a binomial distribution and LINK = LOGIT specifies the logit link that
fits a logistic regression model ***/;
REPEATED SUBJECT=IDNUM / TYPE=UNSTRUCTURED;
/*** The repeated statement invokes the GEE approach. SUBJECT = IDNUM identifies what is
repeated and TYPE = UNSTRUCTURED specifies the structure for the working correlation
matrix. When convergence problems arise, other structures should be considered. ***/;
/*** The following ESTIMATE statements compute the test statistics for the effects named
within single quotes. ***/;
        ESTIMATE 'ADDITIVE FOR A' INTERCEPT 0 COTWIN 0 ZYG 0 D1 0 D2 0  ZCO 1
        D1CO 0 D2CO 0 D1Z 0 D2Z 0 D1ZCO 1 D2ZCO 0;
```

```
ESTIMATE 'COMMON FOR A' INTERCEPT 0 COTWIN 1 ZYG 0 D1 0 D2 0 ZCO 0
D1CO 1 D2CO 0 D1Z 0 D2Z 0 D1ZCO 0 D2ZCO 0;
ESTIMATE 'ADDITIVE FOR B' INTERCEPT 0 COTWIN 0 ZYG 0 D1 0 D2 0 ZCO 1
D1CO 0 D2CO 0 D1Z 0 D2Z 0 D1ZCO 0 D2ZCO 1;
ESTIMATE 'COMMON FOR B' INTERCEPT 0 COTWIN 1 ZYG 0 D1 0 D2 0 ZCO
0D1CO 0 D2CO 1 D1Z 0 D2Z 0 D1ZCO 0 D2ZCO 0;
ESTIMATE 'ADDITIVE FOR LINKED' INTERCEPT 0 COTWIN 0 ZYG 0 D1 0 D2 0
ZCO 1 D1CO 0 D2CO 0 D1Z 0 D2Z 0 D1ZCO 0 D2ZCO 0;
ESTIMATE 'COMMON FOR LINKED' INTERCEPT 0 COTWIN 1 ZYG 0 D1 0 D2 0
ZCO 0 D1CO 0 D2CO 0 D1Z 0 D2Z 0 D1ZCO 0 D2ZCO 0;
```

Responses within the clusters are assumed to be correlated but the between cluster responses are assumed to be statistically independent.

Table I. Cross-twin cross-trait combinations for dichotomous traits A and B

	Twin 2			
Twin 1	$A\,B$	$\bar{A}\,B$	$A\,\bar{B}$	$\bar{A}\,\bar{B}$
$A\,B$	m_{11}	m_{12}	m_{13}	m_{14}
$\bar{A}\,B$	m_{21}	m_{22}	m_{23}	m_{24}
$A\,\bar{B}$	m_{31}	m_{32}	m_{33}	m_{34}
$\bar{A}\,\bar{B}$	m_{41}	m_{42}	m_{43}	m_{44}

Table II. The 2 x 2 summary tables of frequencies and corresponding probabilities under independence for traits A and B and the linked trait (*S*)

a) Trait *A*

	Twin 2	
Twin 1	A	$\overline{A}$
A	$m_{11} + m_{13} + m_{31} + m_{33}$ π_A^2	$m_{12} + m_{14} + m_{32} + m_{34}$ $\pi_A(1 - \pi_A)$
$\overline{A}$	$m_{21} + m_{23} + m_{41} + m_{43}$ $\pi_A(1 - \pi_A)$	$m_{22} + m_{24} + m_{42} + m_{44}$ $(1 - \pi_A)^2$

b) Trait *B*

	Twin 2	
Twin 1	B	$\overline{B}$
B	$m_{11} + m_{12} + m_{21} + m_{22}$ π_B^2	$m_{13} + m_{14} + m_{23} + m_{24}$ $\pi_B(1 - \pi_B)$
$\overline{B}$	$m_{31} + m_{32} + m_{41} + m_{42}$ $\pi_B(1 - \pi_B)$	$m_{33} + m_{34} + m_{43} + m_{44}$ $(1 - \pi_B)^2$

c) Linked trait (L)

	$L_B = 1$	$L_B = 0$
$L_A = 1$	$m_{11} + m_{22} + m_{33} + m_{44}$ $(\pi_A^2 + (1-\pi_A^2))(\pi_B^2 + (1-\pi_B^2))$	$m_{12} + m_{21} + m_{34} + m_{43}$ $2\pi_A(1-\pi_A)(\pi_B^2 + (1-\pi_B)^2)$
$L_A = 0$	$m_{13} + m_{31} + m_{24} + m_{42}$ $2\pi_B(1-\pi_B)(\pi_A^2 + (1-\pi_A)^2)$	$m_{14} + m_{23} + m_{32} + m_{41}$ $4\pi_A\pi_B(1-\pi_A)(1-\pi_B)$

Table III. Variable definition for modeling the linked trait

Twin 1		Twin 2			
Trait A	Trait B	Trait A	Trait B		
Y_1	Y_2	X_1	X_2	L_A	L_B
A	B	A	B	1	1
A	$\overline{B}$	$\overline{A}$	B	0	0
$\overline{A}$	B	A	$\overline{B}$	0	0
$\overline{A}$	$\overline{B}$	$\overline{A}$	$\overline{B}$	1	1
A	B	$\overline{A}$	B	0	1
A	$\overline{B}$	A	B	1	0
$\overline{A}$	B	A	B	0	1
A	B	A	$\overline{B}$	1	0
A	$\overline{B}$	A	$\overline{B}$	1	1
A	B	$\overline{A}$	$\overline{B}$	0	0
$\overline{A}$	B	$\overline{A}$	B	1	1
$\overline{A}$	$\overline{B}$	A	B	0	0
A	$\overline{B}$	$\overline{A}$	$\overline{B}$	0	1
$\overline{A}$	$\overline{B}$	$\overline{A}$	B	1	0
$\overline{A}$	B	$\overline{A}$	$\overline{B}$	1	0
$\overline{A}$	$\overline{B}$	A	$\overline{B}$	0	1

Table IV. Observed frequencies in twin pairs for the traits trouble sleeping and trouble concentrating

Traits in Twin 1		Traits in Twin 2 Trouble Sleeping =Yes Trouble Concentrating =Yes	Trouble Sleeping =No Trouble Concentrating =Yes	Trouble Sleeping =Yes Trouble Concentrating = No	Trouble Sleeping =No Trouble Concentrating =No
Trouble Sleeping	Trouble Concentrating				
			MZ Twins		
Yes	Yes	165	36	123	141
No	Yes	38	15	45	78
Yes	No	116	44	189	252
No	No	111	90	243	787
			DZ Twins		
Yes	Yes	109	29	107	147
No	Yes	39	15	34	68
Yes	No	113	31	142	243
No	No	147	76	216	518

Table V. Twin odds ratios for trouble sleeping, trouble concentrating and the linked trait

Trait		GEE Odds Ratios	se
Trouble Sleeping	MZ	2.776	0.234
	DZ	1.643	0.143
Trouble Concentrating	MZ	2.671	0.266
	DZ	1.667	0.180
Linked trait	MZ	1.655	0.147
	DZ	1.398	0.130

Table VI. Hypothesis tests for additive genetic and common environmental effects for trouble sleeping, trouble concentrating and the linked trait

	GEE	
Effects	Chi-Square	P value
Trouble Sleeping		
Additive Genetics	18.08	0.0001
Common Environment	0.02	0.8863
Trouble Concentrating		
Additive Genetics	10.27	0.0013
Common Environment	0.03	0.8681
Linked trait		
Additive Genetics	1.84	0.1890
Common Environment	0.66	0.4175

Table VII. Observed frequencies in twin pairs for the traits lifetime smoking and alcohol drinking

		Traits in Twin 2			
Traits in Twin 1		Smoking =Yes Drinking =Yes	Smoking =No Drinking = Yes	Smoking = Yes Drinking = No	Smoking =No Drinking =No
Smoking	Drinking	MZ Twins			
Yes	Yes	965	132	15	12
No	Yes	143	308	1	26
Yes	No	10	3	4	3
No	No	15	31	1	43
		DZ Twins			
Yes	Yes	720	145	12	25
No	Yes	145	161	2	14
Yes	No	12	4	5	3
No	No	14	15	0	16

Table VIII. Twin odds ratios for lifetime cigarette smoking, and alcohol drinking and the linked trait

Trait		GEE Odds Ratios	se
Cigarette Smoking	MZ	16.78	2.142
	DZ	5.40	0.73
Alcohol Drinking	MZ	24.83	5.56
	DZ	11.46	3.39
Linked trait	MZ	1.34	.27
	DZ	1.35	0.25

Table IX. Hypothesis tests for additive genetic and common environmental effects for lifetime cigarette smoking, alcohol drinking and the linked trait

	GEE	
Effects	Chi-Square	P value
Cigarette Smoking		
Additive Genetics	37.46	0.0001
Common Environment	3.47	0.0626
Alcohol Drinking		
Additive Genetics	4.02	0.0049
Common Environment	7.58	0.0059
Linked trait		
Additive Genetics	0.01	0.9902
Common Environment	0.51	0.4740